Least-Squares Structuring, Clustering, and Data Processing Issues *†

Boris Mirkin

DIMACS, Rutgers University

96 Frelinghuysen Road

Piscataway, NJ 08854-8018 USA

and Central Economics-Mathematics Institute, Moscow, Russia

Abstract

Approximation structuring clustering is an extension of what is usually called "square-error clustering" onto various cluster structures and data formats. It appears to be not only a mathematical device to support, specify and extend many clustering techniques, but also a framework for mathematical analysis of interrelations among the techniques and their relations to other concepts and problems in data analysis, statistics, machine learning, data compression and decompression, and design and use of multiresolution hierarchies. Based on the results found, a number of methods for solving data processing problems are described.

Contents

1   Introduction

Clustering can be viewed from different perspectives: statistical estimation, optimization, operations research, knowledge domains, etc. We consider clustering as a discipline devoted to revealing combinatorial cluster structures in a set of data about a phenomenon. The major assumption is that a clustering structure is present in the data table in the same format as the data table itself. We consider, among others, traditional entity-to-feature table data format and, also, contingency (summable) and spatial data. The cluster structures under consideration are subset, partition, hierarchy, box (two subsets associated), and bipartition (two partitions associated). The assumption leads us to suppose that any empirical data table can be viewed as a cluster-structure-generated data plus small residuals. The clustering problem, in this setting, is to find out a cluster structure that minimizes residuals scalarized as the sum of squared residual values (the least-squares criterion).

It should be probably emphasized that the usage of the least-squares criterion here much differs of the dominant tradition. Traditionally, least squares is an approach for fitting a model for which a probabilistic distribution is assumed to underlie the data observed. In such a framework, the data has no meaning on its own being considered just a means to identify parameters of the model. The only properties of the least-squares method, the researcher is interested in, are those of effectiveness of the fitting procedure: is it time-consuming or not, whether there is any bias in the estimates, how consistent the estimates are, etc. In contrast to this approach, the least-squares criterion is considered, in this paper, as just a criterion for finding a structure in the data, with no model underlying the criterion. The author doesn't know, for instance, what kind of model can be suggested for Digits data in Table 5 below describing a pattern of presence/absence for segments whose combinations are symbols for the ten numerals. In such a situation, the criterion may become one of many heuristical tools to compute something that has no theoretical meaning. To cope with this, the author suggests theoretically substantiating the criterion not at the input, the model of data, as is usually done, but at the output, the structure found. For instance, there is no good model to justify the use of least squares with discrete or qualitative data (see section 4.2). However, we can see that the averaging of the corresponding zero/one values gives us just those conditional and unconditional frequencies that one would exploit anyway. Moreover, it is proved that this criterion, under the usage suggested, coincides with a known criterion in conceptual clustering, the so-called category utility function (see section 4.4). This, as well as the other findings reported in section

4, may be considered a theoretical justification of the least squares applied to categories. This kind of substantiation based not on a model of the world but rather on similarities proven between seemingly different heuristical approaches, seems at least deserving consideration as a theoretical matter.

Due to the "additive" properties of corresponding mathematical and computational constructions, the square data scatter can be decomposed into two parts, one explained by the cluster structure and the other, unexplained, part (which is equal to the least-squares criterion minimized). This leads us to believe that the data scatter and its explained part must play a major role in cluster analysis. In particular, a number of observations concerning such issues in data processing as machine learning a subset, data mining, mixed feature data analysis, interrelation between classic and conceptual clustering, etc., can be made in terms of this decomposition. Certain implications are found also for less traditional data types such as contingency and spatial data. For spatial data, for instance, due to a representation of data via binary hierarchies, clustering appears explicitly related with such issues in data processing as wavelet-based approximations and quadtrees.

The goal of this paper is to describe the scope and range of data processing issues related to structuring and clustering that could be treated with the least-squares approach. Although most of the author's results have been described elsewhere (see references to the author's work), this text intends to highlight them differently: in the perspective of the data processing issues rather than from the point of view of the methods themselves. The contents of the paper are as follows. In section 2, the data processing issues mentioned above are presented for further treatment. An additive approximation data structuring model is developed in section 3; the model generalizes those in earlier publications. Section 4 considers three popular clustering techniques, K-Means, agglomerative/divisive clustering, and conceptual clustering, in the least-squares framework. Data-scatter based preprocessing is employed as the major facility to process quantitative and qualitative features simultaneously. In section 5, it is shown that some issues in data mining and machine learning can be treated within the framework of the least-squares single cluster clustering. Section 6 is devoted to problems in clustering with relatively nontraditional data type, contingency (or flow) data, characterized by the property that they can be meaningfully summed up across the table. This kind of data seems of great current interest since a contingency table may summarize a really large data set. Two structuring methods are described: box-clustering (revealing the most deviant patterns of interrelation) and aggregation (revealing similar interaction patterns). In section 7, a connection is established between the least-squares hierarchical modelling and two popular image processing concepts, wavelet and quadtree. We indicate some new opportunities emerging due to the fact that the standard "continuity" and "equality" requirements of the latter two concepts can be easily relaxed in the context of hierarchic trees. This may lead to more effective, cluster-based, methods for storage and processing of spatial data.

Specific applications of the general data structuring model involve different data types; we believe that no confusion can occur when we use sometimes the same symbol to denote different things related to different data patterns (as, for instance, when V denotes the set of columns representing features and categories in quantitative presentation of a mixed feature entity-to-feature data table (section 4) and the set of rows involved in a box, $V \times W$, employed in analysis of flow (contingency) data in section 6.

2 A List of Issues in Data Processing

2.1 Machine Learning a Prespecified Subset

The problem of learning a prespecified subset has been extensively considered in the literature on pattern recognition and machine learning. The most popular techniques – discriminant analysis, neural nets, conceptual clustering – all deal mostly with the problem of learning a prespecified subset of the entities. Still each of the approaches has some drawbacks. (Discriminant analysis is well developed only for relatively simple separating surfaces; the neural nets' solutions do not admit simple interpretations; conceptual clustering is oriented to describing, with equal accuracy, both the subset and its complement even if the latter comprises nonhomogeneous entities; etc.) This makes any new strategy a welcome supplement to the existing techniques.

One of the most attractive ideas in machine learning is of finding a distinctive description of a prespecified subset via conjunction of the most important categories (either nominal ones or quantitative intervals or both). For instance, a subset of data entities can be prespecified as, say, articles on finance matters in a given body of articles characterized by their keywords. Then the question is how this subset can be summarized in a compact description involving the keywords in such a way that the description distinctively separates the subset from the other articles. A description like "The Dow Jones index is mentioned more than 3 times and term 'security' also occurs" is good if the articles satisfying it are overwhelmingly concentrated within the finance article subset.

However, finding a distinctive conjunctive description is not an easy task. First, it may require looking at an enormous number of category combinations, and, second, there may be no good descriptions with the given features at all! The question is if any reasonable strategy can be developed to address both of the issues. See section 5.2 for an answer.

2.2 Data Mining as Finding and Describing Interesting Patterns

A recently emerged area of data processing, data mining, is aimed at finding and describing interesting patterns in data.

A nice formulation of what is interesting is this: "discovering the most significant changes in the data from previously measured or normative values" (Fayyad et al. 1996, p. 16). In this treatise, we consider only static data tables so that no "previously measured values" are assumed. The issue is whether this formulation may fit not only in the problem of finding an interesting pattern, but also in the problem of finding an "interesting" description for a pattern prespecified. If yes, issues still remain related to which "normative" values, what measure of change and what thresholds should be employed, and how these may relate to finding distinctive machine learning descriptions. Answers are in section 5.

2.3 Dealing with Mixed Feature Data

The data base records usually are characterized by a set of features (variables) some of which have been measured in quantitative scales while the others are qualitative. Methods for analysis of the records (entities) described in mixed feature space are still an issue.

Consider, for instance, Table 1 where eight masterpieces of Russian literature are presented along with the values of 5 variables, which are: 1) LenSent - Average length of sentences (number of words); 2) LenDial - Average length of dialogues (number of sentences); 3) NChar - Number of principal characters in the novel; 4) InMon - Does the author use internal monologues of the characters or not; 5) Presentat - Principal way of presentation of the subject by the author. The variables 1 to 3 are quantitative, which means that, typically, statements involving quantitative comparisons of their values or quantitative transformations of those, are meaningful. Variable 4 is Boolean (binary); its categories are Yes or No. Variable Presentat is nominal; it has three mutually exclusive categories: Direct - meaning that the author prefers direct descriptions and comments, Behav - the author prefers expressing his ideas through behavior of the characters, and Thought - the subject is shown, mainly, through characters' thoughts.

Table 1: **Masterpieces:** Russian masterpieces of 19th century: the first three by A. Pushkin, the next three by F. Dostoevski, and the last two by L. Tolstoy.

Title	LenSent	LenDial	NChar	InMon	Presentat
Eug.Onegin	15.0	16.6	2	No	Direct
Doubrovski	12.0	9.8	1	No	Behav
Captain's					
Daughter	11.0	10.4	1	No	Behav
Crime and					
Punishment	20.2	202.8	2	Yes	Thought
Idiot	20.9	228.0	4	Yes	Thought
A Raw Youth	29.3	118.6	2	Yes	Thought
War & Peace	23.9	30.2	4	Yes	Direct
A.Karenina	27.2	58.0	5	Yes	Direct

Two major data analysis problems: finding patterns of correlation among the variables and finding patterns of structure in the set of entities, cannot be properly treated without transformation of the data into a quantitative format. In statistics, distribution based methods involving both discrete and continuous variates have been developed only for a limited number of problems. In clustering, the popular approach is to transform the data table into a record-to-record dissimilarity matrix, which is then to be treated by a clustering procedure.

The other natural approach, treating symbolic categories as quantitative dummy variables (see Table 2) is not very popular perhaps because there are no answers yet to the questions of: (1) comparative weighting of those dummy variables against the raw quantitative ones, and (2) meaning of the results of quantitative operations with the dummy variables.

Thus, there is an issue of developing a meaningful strategy for processing mixed feature data as transformed into the format of Table 2. See section 4.2 for an answer (some parts of sections 4 and 5 are also related).

Table 2: Quantitative presentation of the Masterpieces data as an 8 by 7 entity-to-variable/category matrix.

Num	LenSent	LenD	NChar	InMon	Direct	Behav	Thought
1	15.0	16.6	2	0	1	0	0
2	12.0	9.8	1	0	0	1	0
3	11.0	10.4	1	0	0	1	0
4	20.2	202.8	2	1	0	0	1
5	20.9	228.0	4	1	0	0	1
6	29.3	118.6	2	1	0	0	1
7	23.9	30.2	4	1	1	0	0
8	27.2	58.0	5	1	1	0	0

2.4 Classical and Conceptual Clustering: Any Relation?

"Classical" or "traditional" clustering considers the entities to be clustered as points of a geometrical space and formalizes the clusters to be found as "coherent" point groups in the space. With such an approach, the computations do not much depend on the number of the features; however, interpretation of the results may become an issue (see, for instance, Michalski and Stepp, 1992, p. 169). To overcome this, another clustering paradigm, conceptual clustering, has been developed. This, in fact, is based on consideration of the correlations between the features present and classification to be constructed; the total correlation score is measured by such coefficients as "twoing rule" in Breiman et al., 1984, or category utility function (Fisher, 1987). The clustering tree is formed in terms of the feature categories and thus is easy to interpret. Moreover, the calculation does not much depend on the number of entities involved, but it does limit the number of features, because there are difficulties in interpreting clusters when the number of describing categories becomes large.

This makes reasonable the question whether there exists any regular relation between the two approaches. If the answer is "yes", this relation can be employed to combine the results of the two approaches or just to use that approach which is more convenient, in any particular situation. See section 4.4 for an answer.

2.5 Flow Data: Partition or/and Aggregation?

Flow data tables can be distinguished as based on the summability property. Let us take a look at Table 3 (from L. Guttman, 1971, cited by Greenacre, 1988) which cross-tabulates 1554 Israeli adults according to their living places as well as, in some cases, that of their fathers (column items) and "principal worries" (row items). There are 5 column items considered: EUAM - living in Europe or America, IFEA - living in Israel, father living in Europe or America, ASAF - living in Asia or Africa, IFAA- living in Israel, father living in Asia or Africa, IFI - living in Israel, father also living in Israel. The principal worries are: POL, MIL, ECO - political, military and economical situation, respectively; ENR - enlisted relative, SAB - sabotage, MTO - more than one worry, PER - personal economics, OTH - other worries. The columns and the rows of such a matrix correspond to qualitative categories, and its entries represent counts or proportions of the cases fitting both the column and the row categories. In clustering constructions this kind of matrix still has been used

Table 3: **Worries:** The data on cross-classification of 1554 individuals by their worries and origin places.

	EUAM	IFEA	ASAF	IFAA	IFI
POL	118	28	32	6	7
MIL	218	28	97	12	14
ECO	11	2	4	1	1
ENR	104	22	61	8	5
SAB	117	24	70	9	7
MTO	42	6	20	2	0
PER	48	16	104	14	9
OTH	128	52	81	14	12

rather rarely, though it has the obvious advantage of being quite homogeneous. It has a *summability* property: the row or/and column items can be aggregated, according to their meaning, in such a way that the corresponding rows and columns are just summed together. For instance, let us aggregate the columns in Table 3 according to person's living places, thus summing up the columns IFEA, IFAA, and IFI into the aggregate column I (living in Israel) while aggregating their worries into two basic kinds: the worries coming outside their families (OUT=POL+MIL+ECO+SAB+MTO) and inside the families (FAM=ENR+PER); the other worries row OTH remains non-aggregated. The resulting data set:

	EUAM	I	ASAF
OUT	506	147	223
FAM	152	74	165
OTH	128	78	81

Obviously, there can be other sources of summable data than just counting individuals: money or volume or mass flows, for instance. Summability seems quite important in cluster analysis since it makes a natural aggregate representation for any data part related to a cluster. The major clustering constructions, such as K-Means or agglomerative clustering, represent clusters by their averaged, not total, values, which means that the data is actually considered as being just in the entity-to-variable format. There is nothing bad in that. However, to exploit the summability of the data, a technique should be developed to maintain clusters as aggregates, thus combining clustering and aggregating for this kind of data. This is addressed in sections 6.1 and 6.2.

2.6 Category-to-Category Interaction via Contingency Tables

The flow data frequently have the form of an interaction table, as, for instance, in Table 4 reporting results of a psychophysical experiment on confusion between segmented numerals (see Fig. 1) from Keren and Baggen, 1981.

The drawing in Fig. 1 can be transformed into a binary data matrix as presented in Table 5. The seven binary variables correspond to the columns of the data matrix, and the digits, to the rows. The answer "no" is denoted by a missing entry.

7

Table 4: **Confusion:** Keren and Baggen (1981) data on confusion of the segmented numeral digits 0 to 9.

Stimulus	Response									
	1	2	3	4	5	6	7	8	9	0
1	877	7	7	22	4	15	60	0	4	4
2	14	782	47	4	36	47	14	29	7	18
3	29	29	681	7	18	0	40	29	152	15
4	149	22	4	732	4	11	30	7	41	0
5	14	26	43	14	669	79	7	7	126	14
6	25	14	7	11	97	633	4	155	11	43
7	269	4	21	21	7	0	667	0	4	7
8	11	28	28	18	18	70	11	577	67	172
9	25	29	111	46	82	11	21	82	550	43
0	18	4	7	11	7	18	25	71	21	818
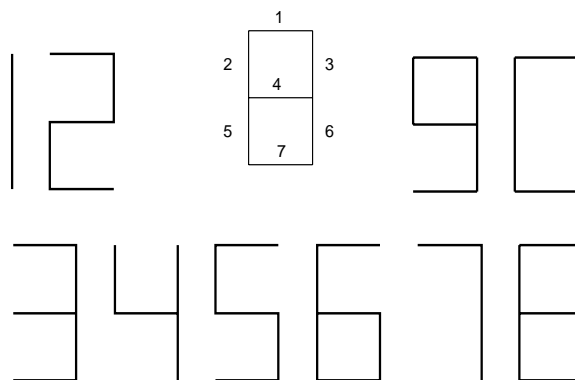
Figure 1: **Digits:** Integer digits presented by segments of the rectangle.

This brings us to the question of aggregating the confusion data table in such a way that each of the aggregate groups consists of entities having similar confusion patterns. An obvious response to this question – calculate a (dis)similarity index between rows (columns) of the Confusion data table and apply a standard clustering algorithm – seems unsatisfactory here because there is no hint which of the numerous (dis)similarity measures should be selected. The issue may be resolved if an aggregation method can be developed to deal with the raw confusion data themselves, not with dissimilarities. Yet another problem concerns interpretation (learning) the eventual confusion classes in terms of the digit segments in Fig. 1 and Table 5: this may suggest an explanation of the confusion patterns in terms of the segments (see section 6.3).

Such an aggregation method, obviously, will have a larger application area to be applied every time when there is a hypothesis that a detailed interaction process may be ruled by an aggregate categorization, as for instance, in analysis of inter-citation data, or international trade data, brand-switching, mobility or input-output industrial data.

Table 5: **Digits:** Segmented numerals presented with seven binary variables corresponding to presence/absence of the corresponding segment in Fig. 1.

Digit	e1	e2	e3	e4	e5	e6	e7
1			1			1	
2	1		1	1	1		1
3	1		1	1		1	1
4		1	1	1		1	
5	1	1		1		1	1
6	1	1		1	1	1	1
7	1		1			1	
8	1	1	1	1	1	1	1
9	1	1	1	1		1	1
0	1	1	1		1	1	1

2.7 Multiresolution Data Processing with Wavelets

Wavelet based multiresolution approximation techniques currently dominate the area of signal and image processing. Such techniques are based on a dilation/translation family of "scale" functions $\chi_{mt}(x) = 2^{m/2}\chi(2^m x - t)$ (defined by a 'simple' function $\chi(x)$ that is zero outside the interval $[0,1]$). For given m, functions $\chi_{mt}(x)$ span a subspace, V_m, of the resolution level m, while complementary functions, the wavelets $\phi_{mt}(x) = 2^{m/2}\phi(2^m x - t)$, span the orthogonal complement, $W_m = V_{m+1} \ominus V_m$ of V_m in V_{m+1}. Perhaps the simplest scale function is the so-called box function whose graph along with the corresponding wavelet function graph (called Haar basis) is shown in Fig. 2. These concepts allow fast data compression/decompression of the signal or image data via level-to-level recalculations of the coefficients of linear representations of the data in the spaces V_m (m=0,1,...). The representations also allow fast and memory-effective approximation of the data (see, for instance, Kay, 1994).
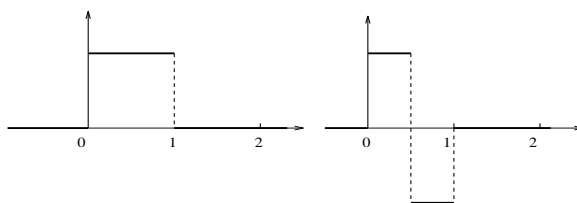
Figure 2: Graphs of the scale and wavelet functions in the Haar basis.

Scaled compression and decompression of a brightness data vector via the Haar basis scale and wavelet dilation/translation functions is presented in Fig. 3. The spaces V_m and W_m correspond to the layers of the trees. The decompositions of the data in spaces V_m, by half-sums of the preceding layer data are shown in the left part, and respective decompositions of the data in the orthogonal subspaces W_m, by the differences of the preceding layer values in the left part, are shown in the right part. Having the overall average, 2, and the right part of Fig. 3, the data in the left part are easy to decompress up to any degree of exactness via subtractions of the right-hand tree values from a left-hand tree value corresponding to their parent. Since in practical calculations the right, wavelet, part of the picture has many zero or almost zero values, layer-to-layer compression/decompression

techniques are very effective.
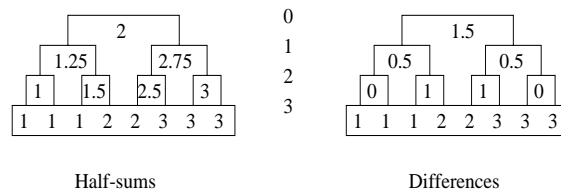
Figure 3: Compression and decompression of the eight-digit data vector with Haar based wavelets.

Actually, multiresolution approximation techniques decompose and maintain spatial data in a specific hierarchical structure, as in Fig. 3, which is called complete: its clusters (the hierarchy nodes) correspond to continuous fragments of the space, and every node is divided into two subnodes of equal sizes. These features may be burdensome, however, when more effective data compression can be achieved by finding and maintaining clusters that may be neither of equal sizes nor spatially restricted to simplest shapes. The question is: can a wavelet-like hierarchy be developed and maintained as just a discrete structure to allow the discontinuities and inequalities? If yes, this potentially may lead to effective data processing techniques based on restructuring computations according to the data patterns (see section 7.1).

2.8   Quadtrees as a Tool for Processing Data

The quadtree is a hierarchical structure for storing image data (see, for example, Samet, 1990). It is designed by sequentially quadrupling the square image portions into four subsquares of equal sizes until the brightness in a subsquare becomes more or less constant. The questions are: can such a structure be utilized in parallel to multiple resolution approximation for fast level-to-level compression/decompression of the data? Moreover, can such a structure be designed in a flexible way so that the conditions of continuity and equality of the 'subsquares' can be relaxed? Potentially, such a modification may lead to more effective methods for image storing and processing (see section 7.2).

3   Approximation Structuring and Clustering

3.1   Additive Structuring Model

The additive structuring model suggests that a data table can be decomposed into the sum of structure-generated tables, each describing an 'ideal' structure supposedly participating in generation of the data. For example, we may think of the Masterpieces data in Table 2 as representing the three different writer styles: those of A. Pushkin (entities 1, 2 and 3), F. Dostoevski (entities 4, 5, and 6), and of L. Tolstoy (entities 7 and 8). It can be further assumed that each of the styles, on average, corresponds to the averaged feature values within the corresponding cluster. These three author styles combined may be regarded as an additive decomposition of the data matrix (up to relatively small residual values), which is presented below (for the sake of space, only three features are shown: LenDial, NChar, and InMon). Moreover, in this example, the integer-valued variables,

as NChar and InMon, keep integers as the cluster values, which is usually not the case within the least squares framework: the Pushkin cluster NChar value might very well be the average, 1.33. This shouldn't embarrass anybody: just a cluster should be considered not as yet one more entity but as a collection, so that 1.33 can be correctly interpreted as a profile value, that is, NChar equals either 1 or 2 within the cluster and the number of 1-cases is as twice as large as the number of 2-cases. Some other examples of additive structural decompositions for various types of data and various types of structures can be found in Mirkin, 1996a, Hubert and Arabie, 1994, 1995.

$$
\begin{pmatrix}
\begin{array}{ccc}
LenD & NChar & InMon \\ \hline
16.6 & 2 & 0 \\
9.8 & 1 & 0 \\
10.4 & 1 & 0 \\ \hline
202.8 & 2 & 1 \\
228.0 & 4 & 1 \\
118.6 & 2 & 1 \\ \hline
30.2 & 4 & 1 \\
58.0 & 5 & 1
\end{array}
\end{pmatrix}
=
\begin{pmatrix}
\begin{array}{ccc}
LenD & NChar & InMon \\ \hline
12.3 & 1 & 0 \\
12.3 & 1 & 0 \\
12.3 & 1 & 0 \\ \hline
183.1 & 3 & 1 \\
183.1 & 3 & 1 \\
183.1 & 3 & 1 \\ \hline
44.1 & 4 & 1 \\
44.1 & 4 & 1
\end{array}
\end{pmatrix}
+
\begin{pmatrix}
\begin{array}{ccc} \hline
4.3 & 1 & 0 \\
-2.5 & 0 & 0 \\
-1.9 & 0 & 0 \\ \hline
19.7 & -1 & 0 \\
44.9 & 1 & 0 \\
-64.5 & -1 & 0 \\ \hline
-13.9 & 0 & 0 \\
13.9 & 1 & 0
\end{array}
\end{pmatrix}
=
$$

$$
\begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}
\begin{pmatrix} 12.3 & 1 & 0 \end{pmatrix}
+
\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}
\begin{pmatrix} 183.1 & 3 & 1 \end{pmatrix}
+
\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}
\begin{pmatrix} 44.1 & 4 & 1 \end{pmatrix}
+
\begin{pmatrix}
\begin{array}{ccc} \hline
4.3 & 1 & 0 \\
-2.5 & 0 & 0 \\
-1.9 & 0 & 0 \\ \hline
19.7 & -1 & 0 \\
44.9 & 1 & 0 \\
-64.5 & -1 & 0 \\ \hline
-13.9 & 0 & 0 \\
13.9 & 1 & 0
\end{array}
\end{pmatrix}
$$

This line of thinking leads us to the following additive structuring model of data, which generalizes the model in Mirkin, 1990.

Let us assume the data is a vector, y, in Euclidean space R^l. The dimensionality l may vary depending on the data format. For instance, $l = N \times (N-1)/2$ when the data is a symmetric $N \times N$ similarity matrix without diagonal entries, or $l = N \times K$ when the data is an entity-to-feature table presented as a $N \times K$ matrix.

Let us assume also that a subset $D \subset R^l$ consists of structures that are considered admissible for the additive representation of the data vector. For instance, when $l = N \times (N - 1)/2$, D may contain all $N \times (N - 1)/2$ binary matrices corresponding to subsets of an N-element set I whose elements label the row/columns of the similarity matrix represented by y. To correspond to a subset, $S \subset I$, such a matrix, $s = (s_{ij})$, $i, j = 1, ..., N$ and $i < j$, can be defined so that $s_{ij} = \alpha$ when both i and j are in S and $s_{ij} = 0$ otherwise. Here, α is supposed to be any positive real; its value may be interpreted as an overall 'intensity' degree of mutual interconnection between entities in S. Similarly, a subset $S \subset I$ (here I is the row set), of the rows of a $l = N \times K$ entity-to-feature matrix, along with the corresponding K-dimensional centroid vector, c, can be represented as a $l = N \times K$ matrix having all its rows corresponding to $i \in S$ equal to c, and all its rows corresponding to $i \notin S$ equal to 0, as in the decomposition of the Masterpieces data above. The set D may consist of all ultrametric or Robinsonian $N \times (N - 1)/2$ matrices (as in Arabie and Hubert, 1994). Some other admissible structures will be discussed in the material below.

It is not assumed here, though, that the admissible structure set D is topologically continuous; the only assumption we need is that D be a cone:

Assumption 3.1 *If $z \in D$, then $\alpha z \in D$, for any real α.*

The additive structuring model then can be defined as the equation

$$y = z_1 + z_2 + ... + z_m + e \tag{1}$$

where $z_1, z_2, ..., z_m$ are to be found in D to minimize the least squares criterion,

$$L_2 = \sum_{i=1}^{l} e_i^2 = \sum_{i=1}^{l} (y_i - z_{i1} - z_{i2} - ... - z_{im})^2 \tag{2}$$

The number m is defined based on prior or posterior considerations.

The model comprises many previous works as its specific cases (see, for instance, Shepard and Arabie, 1979, Chaturvedi and Carroll, 1994, Hubert and Arabie, 1994, 1995, Mirkin, 1990, 1996a, 1996b, 1997a, Mirkin, Arabie and Hubert, 1995). The methods based on the singular-value-decomposition such as principal component analysis also can be considered its specific cases ($l = N \times K$ and D consists of all $N \times K$ matrices whose rank is 1, for the principal component analysis case, see Mirkin, 1990, 1996a).

3.2 Sequential Fitting Strategy

Finding an exact solution to the model in such a general setting seems an unresolvable task. The author has proposed a strategy of sequentially extracting 'structures' $z_1, z_2, ..., z_m$ from 'data' y one-by-one, at each step solving a simpler problem of finding only one structure $z \in D$ by minimizing

$$L_2 = \sum_{i=1}^{l} (y_i - z_i)^2 \tag{3}$$

where y stands for the part of the initial data remaining after the previously found structures have been subtracted from it.

This strategy extends the standard principal component analysis procedure of finding principal component scores and loadings one-by-one and is based on additive formulation of both equation (1) and criterion (2). It is called SEFIT (sequential fitting) or iterative projection in Mirkin, 1990. An exact formulation is this.

Algorithm SEFIT

Step (0) Put the number of iteration, $t = 0$.
Step (1) Add 1 to t and find $z \in D$ (locally) minimizing (3). Put the solution as $z_t \Leftarrow z$;
Step (2) Take residual data

$$y_{iv} \Leftarrow y_{iv} - z_t$$

and go to (1) until Stop-Condition is satisfied.

12

The SEFIT can be considered a sequence of major iterations, t, each involving a sequence of minor iterations for finding a locally optimal solution at step (1) (for any t fixed). It should be noted that D does not need to be the same at all iterations t: the procedure and its properties do not change if D changes from iteration to iteration. The Stop-Condition can be based on a prespecified number of major iterations, m, or on the following decomposition of the data scatter:

$$\sum_{i=1}^{l} y_i^2 = \sum_{t=1}^{m} \sum_{i=1}^{l} z_{it}^2 + \sum_i e_i^2 \tag{4}$$

As usual in statistics, the equation (4) can be thought of as a decomposition of the data scatter into two parts, that explained by the model (the first term in the right part of (4)) and that part not explained (the last term in (4)) which coincides with the criterion (2). The decomposition (4) allows for stopping the process upon achieving a prespecified proportion of the explained part of the data scatter or when contributions of the individual terms, $\sum_{i=1}^{l} z_{it}^2$, become too small.

Usually in statistics a decomposition like (4) holds when all the model constituents, $z_1, z_2, ..., z_m$, are mutually orthogonal. This is not needed here, due to the iterative character of SEFIT and the local optimality of the individual terms as expressed in yet another assumption.

Assumption 3.2 *The local optimality of z_t in step (1) of SEFIT includes the fact that z_t is the best in the axis αz_t in D.*

In the following, for any two vectors $x, y \in R^l$ such that $x = (x_i)$ and $y = (y_i)$ $(i = 1, ..., l)$, their scalar product is denoted as $(x, y) = \sum_{i=1}^{l} x_i y_i$.

Statement 3.3 *Under Assumptions 3.1 and 3.2, decomposition (4) holds for any set of locally optimal vectors $z_1, z_2, ..., z_m$.*

Proof: Let us denote by y_t the residual after iteration t $(t = 1, ..., m)$ so that $y_t = y_{t-1} - z_t$ and $y_0 = y$. Under Assumption 3.2, the vector y_t is orthogonal to z_t. Thus, by Pythagoras's Theorem,

$$(y_{t-1}, y_{t-1}) = (z_t, z_t) + (y_t, y_t).$$

Summing over $t = 1, ..., m$, and noting that $y_0 = y$ and $y_m = e$, we obtain (4). □

Since SEFIT, in general, does not lead to globally minimizing the 'parallel' criterion in (2), a question arises whether or not SEFIT always exhausts y. In general, the answer is no. However, when D is rich enough to satisfy Assumption 3.4, the answer is yes, if step (1) in SEFIT employs a method leading to better results (see Assumption 3.5).

Assumption 3.4 *The set D includes all vectors $u_i = (0, ..., 0, 1, 0, ..., 0) \in R^l$ having a single 1 at the i-th position $(i = 1, ..., l)$.*

The requirement to solutions to the 'local' problems at step (1) of SEFIT can be expressed in terms of the vectors u_i $(i = 1, ..., l)$ in the assumption above.

13

Assumption 3.5 *The locally optimal solutions found at the step (1) of SEFIT are not worse (with regard to criterion (4)) than vectors αu_i, for all real α and $i = 1, ..., l$.*

Assumptions 3.1 and 3.4 guarantee that D contains enough stock to choose from while assumptions 3.2 and 3.5 refer to the method utilized in solving the local problem at Step 1 of SEFIT: the solution found need not be optimal but still better than the trivial αu_is. Under these assumptions we can prove that SEFIT does finally exhaust the data.

Statement 3.6 *If Assumptions 3.1 through 3.5 hold, than $e_{m+1} = y - \sum_{t=1}^{m} z_t$ converges to 0 as m increases.*

Proof: In terms of vectors $e_{m+1} = y - \sum_{t=1}^{m} z_t$, the problem solved in step (1) of SEFIT is that of minimizing (e_{m+1}, e_{m+1}) by selecting $z_m \in D$. According to Assumption 3.2, the solution satisfies:

$$(e_{m+1}, e_{m+1}) = (e_m, e_m) - (z_m, z_m)$$

where $e_1 = y$.

Let e_{km} be the component of e_m with maximal absolute value. Then, $(e_m, e_m)/l \leq e_{km}^2$. On the other hand, since, by Assumption 3.5, z_m is not worse than any αu_k, we have $(z_m, z_m) \geq y_{km}^2$ because the optimal element of the axis αu_k is equal to $(e_m, u_k)^2/(u_k, u_k) = y_{km}^2$.

Thus, we have

$$(e_{m+1}, e_{m+1}) = (e_m, e_m) - (z_m, z_m) \leq (e_m, e_m)(1 - 1/l)$$

This implies that $(e_{m+1}, e_{m+1}) \leq (y, y)(1 - 1/l)^m$, where $(1 - 1/l)^m$ converges to 0 for increasing m, which proves the statement. $\square$

SEFIT can be modified. One idea is to continue calculations after all m structural components, z_t, have been found by one-by-one updating the z_t with regard to the residual vector $y_t = y - \sum_{s \neq t} z_s$ by minimizing the squared difference (4) for $y = y_t$. Similar ideas have been elaborated in Chaturvedi and Carroll, 1994 and Hubert and Arabie, 1994. A somewhat simpler idea is to recalculate just the norms of z_t, which can be done as follows. Suppose that we denote by P_t the orthogonal projection operator onto the space of the first t vectors $z_1, z_2, ..., z_t$, and on each t-th step let $y - P_t y$ be the residual vector and not $e_m - z_m$. The decomposition in (4) is now lost, but the process converges in a finite number of steps, as follows from the Statement 3.7.

Statement 3.7 *If Assumptions 3.1 through 3.5 hold, the vector z_t found on the t-th iteration of the SEFIT method with $y - P_t y$ used as the residual (for every t) is linearly independent of the vectors $z_1, ..., z_{t-1}$.*

The proof of this is omitted; it closely follows the proof of Theorem 3 in Mirkin, 1990.

4   Least-Squares Criteria in K-Means, Hierarchical and Conceptual Clustering

4.1   Partitioning Model and K-Means

The data vector here is an entity-to-feature data table which is a rectangular array having the rows corresponding to entities and the columns corresponding to features or their categories, with the entries coding values of the features at the entities. Three types of features, quantitative, binary (Boolean) and nominal, encoded as in Table 2, are to be maintained here.

The originally encoded data matrix will be denoted by $X = (x_{iv})$ where $i \in I$ are entities and $v \in V$ are features/categories corresponding to columns. This data is preprocessed into matrix $Y = (y_{iv})$ with the standard preliminary transformation (standardization) so that

$$y_{iv} = \frac{x_{iv} - a_v}{b_v}, \ i \in I, \ v \in V \tag{5}$$

which is a standard interval scale transformation assuming change of both the scale factor (dividing by b) and the origin (adding of a) in the original column x_v. Choice of a and b as well as their meaning for categories will be discussed below after introducing the bilinear clustering model.

In the partitioning model, set D is comprised of additive cluster structures as shown on p. 11. Each additive cluster structure is a set of m clusters, any cluster t, $t = 1, ..., m$, being defined with two objects: 1) its membership function $s_t = (s_{it}), i \in I$, where s_{it} is 0 or 1 characterizing thus a cluster set $S_t = \{i \in I : s_{it} = 1\}$, 2) its standard point, or centroid vector, $c_t = (c_{tv}), v \in V$, to be combined in an $N \times |V|$ cluster-type matrix with elements $\sum_{t=1}^{m} c_{tv} s_{it}$. The model (1) becomes

$$y_{iv} = \sum_{t=1}^{m} c_{tv} s_{it} + e_{iv} \tag{6}$$

The least-squares criterion in this case is equal to

$$\sum_{i,v} e_{iv}^2 = \sum_{t=1}^{m} \sum_{i \in S_t} \sum_{v} (y_{iv} - c_{tv})^2 = \sum_{t=1}^{m} \sum_{i \in S_t} d^2(y_i, c_t) \tag{7}$$

where $d^2(y_i, c_t)$ is Euclidean distance squared between i-th row of Y and centroid c_t.

This criterion is well-known in cluster analysis as the "square error clustering" criterion (Jain and Dubes, 1988). It is well known also, that the parallel K-Means partitioning method is the method of alternating minimization for this criterion: given centroids c_t, $t = 1, ..., m$, the minimal distance rule assigns the entities optimally to the clusters, and, given the memberships, the optimal centroids are the gravity centers.

4.2   The Scatter-Based Standardization and Interpretation Principles

The decomposition (4), in the partitioning model, becomes:

$$\sum_{i \in I} \sum_{v \in V} y_{iv}^2 = \sum_{t=1}^{m} \sum_{v \in V} c_{tv}^2 |S_t| + \sum_{i \in I} \sum_{v \in V} e_{iv}^2, \tag{8}$$

Usually the equation in (8) is interpreted in terms of analysis of variance. In cluster analysis, interpretation of (8) in terms of the contributions to data scatter seems more helpful. Equation (8) decomposes the data scatter into explained and unexplained parts due to the clustering model; moreover, the unexplained part is nothing but the minimized criterion of the model. This is why the present author considers the data scatter as the base for choosing the data standardization parameters in (5).

All the variables should be standardized so that their contributions to the data scatter reflect their relative weights. If the data analyst has no weighting of the variables to suggest, which is a typical situation, the variables should be considered as having equal weights and standardized in such a way that their contributions become equal to each other. This principle of equal contribution makes meaningful comparison of the variables by their contributions to the explained (or unexplained) part of the data scatter. Such comparison may reveal the most contributing, thus salient, variables and categories. The principle should be considered as an adequate formalization of the requirement of equal weight of the variables in numerical taxonomy (Sneath and Sokal, 1973). Usually, in cluster analysis, this requirement is treated in much more vague terms of the between-entity distances.

The choice of parameter a_v does not affect the partitioning model (6), however when the model is set forth in a sequential way with the "component" axes z_t identified not simultaneously, but one-by-one as suggested in the next section, the solution heavily depends on the origin of the variable/category space. To adjust to this kind of principal/correspondence-analysis-like methods, let us postulate an analogue to the law of minimum moment of inertia in mechanics: the origin of the variable space should be a minimizer of the data scatter.

The two scatter-based principles lead to unambigous definitions for the parameters a_v and b_v. For quantitative features, they lead to the usual z-score standardization rule: the origin is the grand mean while the standard deviation is the scale factor, to make the contribution of the feature equal N.

In the case of binary categories, the average of a category $v \in V$ column vector is equal, obviously, to the frequency of the category in I, p_v. To satisfy the principle of equal contribution with $a_v = p_v$, the scale factor can be taken as $b_v = \sqrt{(\#k - 1)p_v}$ where $\#k$ is the number of categories, v, in the nominal variable k. This makes the contribution of the variable k equal to 1. If a category represents just a Boolean variable, the scale factor takes the usual form, $b_v = \sqrt{p_v(1 - p_v)}$.

Having in mind these standardization rules, let us explore the meaning of the contribution of a feature-cluster pair (v, t) to the explained part of the data scatter, which is $c_{tv}^2 |S_t|$. It is proportional to the cluster cardinality and to the squared distance from the grand mean of the variable to its mean (standard value) within the cluster. The contribution of an entity-cluster pair can be evaluated as the scalar product, (y_i, c_t), because $c_{tv}^2 |S_t| = (\sum_{i \in S_t} y_{iv}/|S_t|)c_{tv}|S_t| = \sum_{i \in S_t} y_{iv}c_{tv}$.

To analyze the contributions of nominal variables and their categories to the scatter part explained via cluster partition, let us denote by p_{vt} the proportion of the entities simultaneously having category v and belonging to cluster S_t. Then, for any category v standardized by formula (5), its mean within cluster S_t is equal to $c_{tv} = (p_{vt} - p_t a_v)/(p_t b_v)$. The contribution of a category-cluster pair (v, t) to the explained part of the data scatter is equal to

$$s(v, t) = c_{tv}^2 |S_t| = N(p_{vt} - p_t a_v)^2/(p_t b_v^2), \tag{9}$$

which can be considered a measure of association between category v and cluster t. In particular,

$$s(v, t) = N(p_{vt} - p_t p_v)^2 / (p_t p_v),$$

when $b_v = \sqrt{p_v}$, or,

$$s(v, t) = N(p_{vt} - p_t p_v)^2 / p_t$$

when $b_v = 1$, etc.

These values should be employed for interpretation of comparative salience of features to clusters in analysis of least-squares (or square error) clustering results.

Applied to the Masterpieces data with the number 3 clusters prespecified, K-Means gives the three author clusters. The interpreting coefficients are given in the following Table 6:

Table 6: Cluster structure of the Masterpiece data; in any cluster, the averages of the variables in real and standardized scales are shown in the first and second rows; the third row contains the feature-to-cluster contributions, the contributions expressed in percent are in the fourth row.

Cluster	LenS	LenD	NChar	InMon	Dire	Beha	Tho	Total
	12.67	12.27	1.33	0	0.33	0.67	0	
Pushkin	-1.14	-0.87	-0.62	-1.29	-0.05	0.59	-0.43	
	3.91	2.26	1.16	5.00	0.01	1.04	0.56	13.94
%	28.06	16.22	8.30	35.86	0.04	7.47	4.02	34.86
	23.47	183.13	2.67	1	0	0	1	
Dostoevski	0.55	1.19	0.03	0.775	-0.43	-0.35	0.72	
	0.92	4.25	0.00	1.80	0.56	0.38	1.56	9.48
%	9.71	44.86	0.03	19.01	5.92	3.95	16.48	23.71
	25.55	44.10	4.5	1	1	0	0	
Tolstoy	0.88	-0.48	1.33	0.775	0.72	-0.35	-0.43	
	1.45	0.47	3.54	1.20	1.04	0.25	0.37	8.44
%	17.24	5.57	42.01	14.25	12.35	2.96	4.43	21.09
Total	6.29	6.99	4.70	8.00	1.61	1.67	2.50	31.86
Total,%	15.72	17.46	11.76	20.00	4.02	4.16	6.25	79.66

Table 6 shows that the three clusters count for almost 80% of the data scatter which is equal to $8 \times 5 = 40$. Among the variables, InMon is an obvious leader contributing all its 20% initial weight to the cluster structure. This occurs because the variable is constant in each of the clusters. The contribution of another qualitative feature, Presentat, is only $4.02 + 4.16 + 6.25 = 14.43\%$ of the data scatter, because it is not constant for Pushkin's novels. On the other hand, this variable differentiates between Tolstoy and Dostoevski very clearly, and its category Thought is characteristic for Dostoevski. Why does category Thought not give higher scores? Because, in this example, we don't consider the categories as independently meaningful elements: it is all three, not each, of them that get the weight of a variable, which is equal to $N = 8$ (under standardization applied). Thus, for a particular category to get a higher score, it should be standardized (that is, weighted) differently.

4.3 The Hierarchical Clustering Model and Ward/Edwards/Cavalli-Sforza Criterion

To discuss hierarchical clustering, we consider a binary hierarchy as a set of subsets $S_W = \{S_w : S_w \subseteq I, w \in W\}$ called clusters containing either all singletons (in the case of agglomerative clustering) or I (in the case of divisive clustering, which will be the only one considered here) so that the clusters S_w, $w \in W$, are nested and every non-terminal cluster S_w, $w \in W$, is a union of its two children clusters $S_{w1}, S_{w2} \in S_W$. The terminal clusters have no children.

For any non-terminal cluster $S_w = S_{w1} \cup S_{w2}$ $(w, w1, w2 \in W)$ of S_W, its three-valued *nest indicator function* ϕ_w is defined as follows:

$$\phi_{iw} = \begin{cases} a_w & if & i \in S_{w1} \\ -b_w & if & i \in S_{w2} \\ 0 & if & i \notin S_w \end{cases} \tag{10}$$

where the values a_w and b_w are selected to satisfy the following two conditions: (1) vector ϕ_w is centered; (2) vector ϕ_w has its norm equal to 1. It is easy to see that

$$a_w = \sqrt{\frac{n_{w2}}{n_{w1} n_w}}, \text{ and } b_w = \sqrt{\frac{n_{w1}}{n_{w2} n_w}} \tag{11}$$

where n_w, n_{w1}, and n_{w2} are cardinalities of S_w and its two children, S_{w1} and S_{w2}, respectively.

It turns out that the vectors ϕ_w are mutually orthogonal, $(\phi_w, \phi_{w'}) = 0$, which is trivial when $S_w \cap S_{w'} = \emptyset$ and also true when $S_w \cap S_{w'} \neq \emptyset$ since in the latter case one of the clusters is a part of the other and, thus, its components are non-zero when the other vector's components are constant. Therefore, set $\{\phi_w : w \in W\}$ is an ortho-normal basis of a subspace of all N-dimensional centered vectors, whose components belonging to the same terminal cluster coincide. When all singletons (as well as I) are a part of a binary hierarchy, this means that the subspace, actually, is the $(N-1)$-dimensional space of all N-dimensional centered vectors.

Let $M \leq N$ be the number of terminal nodes in a binary hierarchy S_w, $w \in W$. Then, any data matrix Y whose columns have been centered (so that the sum of components of every column in Y is zero) can be decomposed as follows:

$$Y = \Phi C + E \tag{12}$$

where $\Phi = (\phi_{iw})$ is the $N \times (M-1)$ matrix of the values of the nest indicator functions in (10) and $C = (c_{wv})$ is an $(N-1) \times |V|$ matrix the coefficients of which can be found by minimizing the least-squares criterion, $(E, E) = \sum_{i,v} e_{iv}^2$.

Thus we are again in the realm of the data structuring model, where z_t is a matrix with entries $(\phi_{it} c_{tv})$. This time D certainly depends on t because the definition of ϕ involves a cluster and its split.

The least-squares optimal entries of matrix C expressed through the data are not difficult to find:

$$c_{wk} = \sqrt{\frac{n_{w1} n_{w2}}{n_w}} (y_{w1v} - y_{w2v}) = \sqrt{\frac{n_{w1} n_w}{n_{w2}}} (y_{w1v} - y_{wv}), \tag{13}$$

18

where y_{wv}, y_{w1v} and y_{w2v} are the averages of the variable/category $v \in V$ in S_w, S_{w1} and S_{w2}, respectively. By analogy with the factor loads in principal component analysis, the entries of C can be referred to as cluster loads.

Let us denote by y_w the m-dimensional vector of the averages of the variables in a subset S_w, $w \in W$. The equality in (13) implies that the norm of vector $c_w = (c_{wv})$ can be expressed as

$$\mu_w = \sqrt{\frac{n_{w1}n_{w2}}{n_w}} d(y_{w1}, y_{w2}) \tag{14}$$

where $d(x, y)$ is the Euclidean distance between vectors x, y. The value μ_w is positive if $x \neq y$, and zero if $x = y$. It is an analogue of the singular value in the decomposition (12) considered as an analogue of the singular-value decomposition.

Because of mutual orthogonality of the nest indicator functions, the data scatter decomposition holds as soon as C is defined according to (13):

$$\sum_{i \in I, v \in V} y_{iv}^2 = \sum_{t=1}^{m} \mu_t^2 + \sum_{i \in I, v \in V} e_{iv}^2 \tag{15}$$

so that finding an optimal m-column Φ requires maximizing $\sum_{t=1}^{m} \mu_t^2$.

With SEFIT, at its iteration w, the criterion to maximize becomes

$$\mu_w^2 = \frac{n_{w1}n_{w2}}{n_w} d^2(y_{w1}, y_{w2}), \tag{16}$$

which was used in Ward's, 1963, agglomerative clustering. The same expression was employed by Edwards and Cavalli-Sforza, 1965, for divisive clustering, to be maximized by splitting a cluster S_w into S_{w1} and S_{w2}. The step of taking residual data in SEFIT can be skipped here since it doesn't affect the results, as is not difficult to prove. The task of maximizing criterion (16) is not too hard (it requires enumerating not more than $N^{|V|}$ hyperplanes separating classes in two-class partitions, see Bock, 1974); moreover the standard K-Means method (with two clusters) can be applied as an alternating maximization technique since criterion (16) is equivalent to the least-squares clustering criterion.

The data structuring model employing (16) thus does not provide many new insights in the algorithms, though it may be considered a model-based substantiation of the known principles. However, another form of (16),

$$\mu_w^2 = \frac{N_w N_{w1}}{N_{w2}} d^2(y_{w1}, y_w) \tag{17}$$

may imply a different splitting algorithm since the center y_w of S_w does not vary in the splitting process based on (17) (see Mirkin, 1997a).

The model leads to a number of interpretation aids concerning a binary hierarchy (especially when it is resolved, that is, contains both I and all singletons). One of them is the decomposition (12) of the centered data entries according to hierarchy clusters. Another one is that

$$Y^T Y = C^T C \tag{18}$$

when the hierarchy is resolved. This equation provides for both, the feature variances and covariances decomposed via cluster contributions.
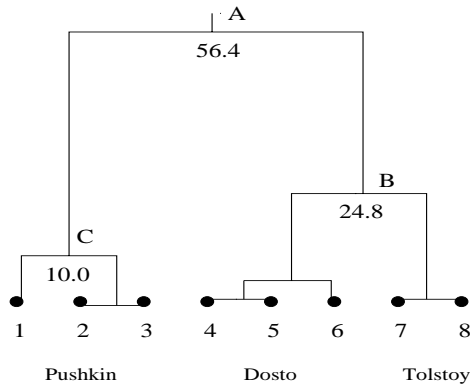
Figure 4: Least-squares divisive clustering results for the Masterpieces data. The vertical scale shows contributions of the splits to the data scatter; the contributions of splits 2-3, 4-5, 7-8 cannot be distinguished at the current drawing scale.

The sequential splitting SEFIT process applied to the Masterpieces data leads to the hierarchy presented in Fig. 4. The major splits A, B, C are labeled by their contributions to the data scatter.

The correlation between LenDial and NChar decomposed via splits A, B, C (according to (18)) is:

$$\begin{array}{ccccccccc} & & Total & & A & & B & & C \\ Corr(LD, NC) & = & 0.27 & = & 0.48 & - & 0.33 & + & 0.07 & +... \end{array}$$

Its large value at split A may be attributed to the fact that the left part (Pushkin cluster) has both, LD and NC, smaller than the right cluster. The negative value in split B says that there is a negative correlation between LD and NC when Tolstoy (moderate LD and large NC) is compared to Dostoevski (large LD and moderate NC).

Decomposition of the value LD for Anna Karenina (entity 8 in Table 2), which is -26.3 after subtracting the grand mean, 84.3, through the clusters (A and B) it belongs to, is

$$x_{A.K.,LD} = -26.3 = 43.22 - 83.42 + 13.90$$

where its increase to 43.22 in B is due to the difference between LD in the left and in the right parts of the tree, and its fall by -83.42 relates to the difference between two last clusters. The last term, 13.90, is the "individual" part of the entry.

4.4 The Least-Squares Criterion and Conceptual Clustering

Conceptual clustering is a discipline related to constructing partitions, starting from the entire set I, by sequentially dividing current clusters by single features. Actually, at each step of construction of a classification tree, the following problems are to be solved:

1. Which class (node of the tree) and by which variable to split?

2. When to stop splitting?

3. How to prune/aggregate the tree if it becomes too large?

In this paper, we concentrate only on item 1 from this list, that is, on defining a goodness-of-split criterion which must depend on the learning task solved by the classification tree. When the tree is for learning the data features, we may check the least-squares partitioning model to provide us with a criterion based on the part of the data scatter explained by the partition.

Since all the contributions are summed up in the decomposition (8), we may consider the contribution of each feature to the explained part separately.

Let us take a nominal variable, k, which is presented in the data by the set of its categories v. The joint contribution of k and the set of the clusters S_t to the scatter of the data is equal to $F(k, S) = \sum_t \sum_{v \in k} s(v, t)$ which is

$$F(k, S) = N \sum_{t=1}^{m} \sum_{v \in k} \frac{(p_{vt} - p_t a_v)^2}{p_t b_v^2} \tag{19}$$

by (9). Substituting the appropriate values of $a_v = p_v$ and b_v, we arrive at the following. For criterion L_2, the contribution of a nominal variable $k \in K$ to that part of the square scatter of the square standardized data which is explained by the (sought or found or expert-given) cluster partition $S = \{S_1, ..., S_m\}$, is equal to

$$\Delta(S/k) = N \sum_{v \in k} \sum_{t=1}^{m} \frac{(p_{vt} - p_v p_t)^2}{p_t} \tag{20}$$

when $b_v = 1$ (no normalizing), or

$$M(S/k) = \frac{N}{\#k - 1} \sum_{v \in k} \sum_{t=1}^{m} \frac{(p_{vt} - p_v p_t)^2}{p_v p_t} \tag{21}$$

when $b_v = \sqrt{p_v(\#k - 1)}$.

These coefficients relate to well known indices of contingency between nominal variables: $M(S/k)$ is a normalized version of the Pearson chi-squared coefficient, and $\Delta(R/k)$ is proportional to the coefficient of reduction of the error of proportional prediction. Thus, the statistical contingency coefficients appear to be contributions to the data scatter, and, moreover, the method of data standardization determines which of the coefficients is produced as the contribution-to-scatter.

The contribution of a quantitative variable to the explained part of the L_2 data scatter is also meaningful. When the variable k is standardized, it is exactly $N\eta^2(k, S)$ where $\eta^2(k, S)$ is the so-called correlation ratio (squared).

These observations lead to the following.

Statement 4.1 *The least-squares partitioning model requires, at each step of the splitting process, maximization of the summary correlation between the partition produced and the variables, $\sum_k \rho(S, k)$, where $\rho(S, k)$ is either the correlation ratio (when k is quantitative) or, when k is nominal, $M(S/k)$ (21) or $\Delta(S/k)$ (20) (depending on whether the categories have been normalized or not).*

21

The statement gives a simple splitting rule for the case when the variables can be both quantitative and categorical. Let us compare this rule with two popular criteria developed for the case when all features are nominal. These criteria are:

1. *Twoing Rule* (Breiman et al., 1984) applied when the split of S is made into two subclasses, S_1 and S_2, only:
$$tw(y, S_1, S_2) = \frac{p_1 p_2}{4} [\sum_u |p(u/S_1) - p(u/S_2)|]^2 \tag{22}$$
   where u are the categories of y.

2. *Category Utility Function* (Fisher, 1987) applied when there is a set of categorical variables Y and the split is made into any number T of subclasses S_t, $t = 1, ..., T$:

$$CU(Y, \{S_t\}) = \sum_{y \in Y} [\sum_{u_y} \sum_t \frac{p_{u_y t}^2}{p_t} - \sum_{u_y} p_{u_y}^2]/T \tag{23}$$

   where u_y is a category of a categorical variable $y \in Y$.

It is easy to see that

Statement 4.2 *The category utility function in (23) is nothing but the sum of Δ coefficients (20): $CU(Y, \{S_t\}) = \sum_{y \in Y} \Delta(S/y)$; that is, it is the explained part of the squared data scatter according to the approximational partitioning model when all variables are presented by nonnormalized categories.*

A conceptual tree derived with the summary correlation function from Statement 4.1 is presented in Fig. 5
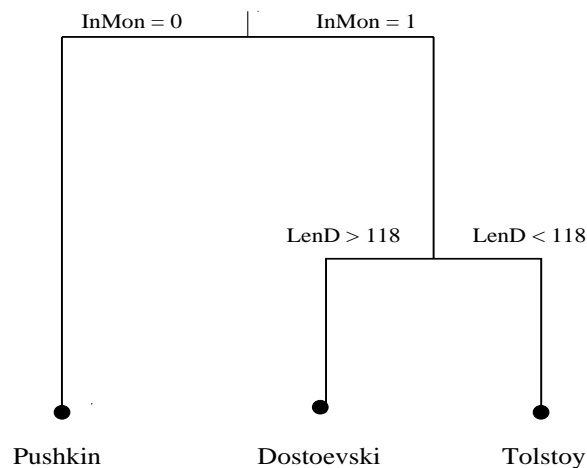

Figure 5: A conceptual tree for the Masterpieces data.

The other, twoing function cannot be so easily interpreted in terms of the partitioning model. However, it has a good match in terms of the hierarchical structuring model. Indeed, criterion μ_w^2

in (16) for splitting in least-squares hierarchical clustering, applied to nominal variables, is equal to

$$\mu_w^2 = n_w p_1 p_2 \sum_u |p(u/S_1) - p(u/S_2)|^2 \tag{24}$$

which much resembles the twoing rule criterion but has a geometrical meaning as well.

The contents of this subsection show that, actually, the difference between conceptual and classical clustering should not be overemphasized; conceptual clustering can be considered as just another local search procedure for optimization of the same or similar criteria as those used in classical clustering.

5 The Most Deviant Patterns: Finding and Describing Single Clusters

5.1 Single Cluster Model and SCC method

Let us consider an approximation structuring model with a single cluster defined by its standard point $c = (c_v)$, $v \in V$, and binary membership vector, $s = (s_i)$, $i \in I$ (both of them may be unknown):

$$y_{iv} = c_v s_i + e_{iv}, \ i \in I, \ v \in V, \tag{25}$$

This model, along with the least-squares criterion,

$$L^2(c, s) = \sum_{i \in I} \sum_{v \in V} (y_{iv} - c_v s_i)^2, \tag{26}$$

can be considered as that of step (1) in SEFIT applied to the m-cluster model in (6) (this time, clusters may be overlapping). This SEFIT-based strategy has been referred as the principal cluster analysis method in Mirkin, 1990, 1997a, since it closely follows the line of computations in the singular-value-decomposition based procedure in the method of principal component analysis.

However, the single clustering model has a meaning on its own. The decomposition (4) of the data scatter here is:

$$\sum_{i \in I} \sum_{v \in V} y_{iv}^2 = d^2(c_t, 0)|S_t| + \sum_{i \in I - S_t} d^2(y_i, 0) + \sum_{i \in S_t} d^2(y_i, c_t) \tag{27}$$

where $d^2(x, y)$ is the Euclidean distance squared between x and y.

This shows that the cluster to determine is that most distant from the origin (which is the grand mean when the data have been standardized preliminarily): its contribution to be maximized is the distance, $d^2(c_t, 0)$, weighted by the cluster's cardinality. This can be considered a model-based explication of the intuitive notion of 'interestingness' (as quoted above from Fayyad et al., 1996). It is the grand mean which is considered here a 'normative value', and it is the Euclidean distance squared between the cluster's gravity center and the grand mean, which measures the deviation. The cluster itself can be considered an explication of the concept of 'interesting pattern'.

To find a cluster based on the model, the other two terms on the right can be exploited as the criterion to minimize. This can be done with a K-Means-like algorithm starting with c equal to

the most distant (from 0) entity point, y_i, ($i \in I$) and then reiterating the two following steps: (a) updating the cluster S as the set of those entity points whose distance to c is smaller than to 0; (b) updating the center c by computing the gravity center of the subset S found on step (a). The process stops when S does not vary anymore. This alternation minimization algorithm is referred to as Separate-and-Conquer Clustering algorithm (SCC) in Mirkin, 1998 (following a earlier suggestion in Pagallo and Haussler, 1990).

5.2 The Contribution Weights and Their Uses for Machine Learning and Data Mining

The decomposition in (27), as well as that in (8), leads us to cluster-specific contribution weights of the features, $c_v^2|S|$; each is proportional to the squared difference between within-cluster mean and grand mean of the corresponding feature. The expression in (9) applies when v is a category. Loosely speaking, the farther c_v is from zero (which is the grand mean here) the more separated is the cluster from the other entities along the "axis" of feature/category v. In terms of Fayyad et al., 1996, this measures the 'degree of interestingness' of the feature v in the cluster with regard to its 'normative' value. It is important to note that this holds for both kinds of situations, when the cluster is (to be) found and when it is prespecified (by a supervisor).

Based on Table 6, it is not difficult to find the most contributing variables/categories for each of the three clusters. This can be employed for finding distinctive logical descriptions of the clusters. For instance, cluster Pushkin can be distinctively described by the fact that InMon (relative contribution is 35.9%, the maximum) is 0 at this cluster. Cluster Dostoevski can be distinctively described by the statement that LenD (the relative contribution is 44.9%, the maximum) is greater than 118. However, it is not that easy for cluster Tolstoy: the most contributing variable NChar (42.0%) does not describe the cluster distinctively with the statement "NChar $\geq$ 4" since a novel by Dostoevski also has NChar=4. Thus, we need to add another feature to the statement to have no false positives. Adding of the next most contributing variable, LenSent (17.2%), does not improve the description since its range in Tolstoy cluster (from 23.9 to 27.2) falls within its range in Dostoevski cluster (from 20.2 to 29.3). Similarly, the next contributing feature, InMon (14.2%) fails to improve the situation. It is the next category, Direct (12.4%), which helps to separate the cluster Tolstoy that thus is described distinctively by the following statement: "NChar $\geq$ 4 & Presentat=Direct".

In general, there is no straightforward relation between the contribution weight of a variable and its "distinctiveness" in logical description of the cluster since the former is a "soft" statistical concept and the latter is quite a rigid one. However, the contribution weight can be employed as a heuristic in finding distinctive descriptions or, at least, in improving the quality of logical description of clusters (Mirkin, 1998).

Let us limit ourselves to conjunctive descriptions of a prespecified cluster S having their conjunctive terms of the form "category $v = A$" or "within-cluster range of quantitative variable v is contained in the interval $[a, b]$". Such a conjunctive description geometrically corresponds to a multidimensional rectangle in the subspace of the features occurred in the description. The degree of distinctiveness of a conjunctive description W can be characterized by the proportions of false positives and false negatives.

Let us assume, for the sake of simplicity, that only quantitative variables occur in the data. This

makes the number of false negatives equal to zero since all within cluster entities will be covered by every term (which is just the within-cluster range).

To minimize the number of false positives, a local search procedure can be developed by formalizing what we have done for describing Masterpieces clusters above. According to this procedure, a number of within-cluster-range-based terms is to be initially collected into a logical conjunction (first phase), after which redundant terms are excluded one-by-one (second phase). The first phase goes along the contribution weight ordering, starting with the empty set of conjunctive terms. Any particular feature is considered according to the ordering to decide whether or not it should be included in the conjunction. It is included only if this decreases the number of false positives. The process stops when there are no features left in the ordering or when the number of false positives becomes zero (or any other prespecified threshold value). The second phase goes in the opposite direction along the terms collected at the first phase to decide whether the single term considered can be removed from the collection or not. It is removed if its removal does not change the number of false positives. This procedure has been described in detail in Mirkin, 1998, as the algorithm of Approximate Conjunctive Concept Learning (ACCL).

Let us apply ACCL to the data in Table 5 to find conceptual descriptions of the Digit classes found by the Confusion table. The four-class partition of the integer digits in the hierarchy of Fig. 9 is $S = \{\{1, 4, 7\}, \{3, 5, 9\}, \{6, 8, 0\}, \{2\}\}$. In the Digit data table , the most contributing variables to the clusters are e7 and e1 (cluster 1), e5 and e7 (cluster 2), e5 and e2 (cluster 3), and e6 (cluster 4). It appears, the four clusters can be described, without errors, by the conjunctive concepts involving the most contributing features: e7=0, e5=0 & e7=1, e5=1 & e2=1, and e6=0, respectively. Perhaps, this can be interpreted as an indication of the most confusing digit segments.

This algorithm performs rather well when the classes are located in different zones of the original feature space. The method works poorly in the domains like the well-known Fisher-Anderson Iris data (see, for instance, Mirkin 1996a and Fig. 6) where classes are intermingled in the feature space so that a class cannot be separated into that box-like cylinder volume which corresponds to an ACCL output conjunction.
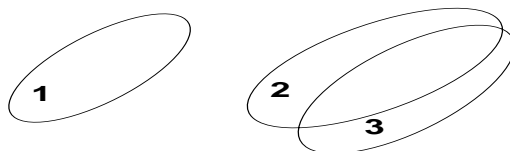

Figure 6: The structure of Iris data set on the plane of two first principal components (150 specimens belonging to 3 genera and described by 4 non-specifying variables).

In the Iris data set, there are four variables, $w1$ to $w4$. The three predefined classes can be described by the following concepts found with algorithm ACCL: $1 \leq w3 \leq 1.9$ (class 1, the number of false positives, FP, is 0), $3.0 \leq w3 \leq 5.1$ & $1.0 \leq w4 \leq 1.8$ (class 2, FP=8), and $1.4 \leq w4 \leq 2.5$ & $4.5 \leq w3 \leq 6.9$ (class 3, FP=18). The errors of the two latter conjunctions cannot be reduced by adding other variables' ranges. Large number of false positives for two of the Iris classes corresponds to their spatial structure in Fig. 6.

However, the method's performance can be improved by transforming and combining the variables (as in Wnek and Michalski, 1994). To do this, denote by A the set of original variables and B a set of ACCL selected variables. Denote by S(A,B) the set of all pair-wise products, xy, ratios, x/y and y/x, sums, $x + y$, and differences, $x - y$, for all $x \in$ A and $y \in$ B. Then iteratively perform ACCL on A, S(A,B), S(A, B(S(A,B)), etc. Such a process usually leads to drastic reduction of the number of false positives in the ACCL results.

Let us take, for example, B= $\{w3, w4\}$, compute S(A,B) and apply ACCL to the set of combined variables. The resulting conjunctions are: $1.18 \leq w1/w3 \leq 1.70$ & $3.30 \leq w3 * w4 \leq 8.64$ (class 2, 4 false positives) and $7.50 \leq w3 * w4 \leq 15.87$ & $1.80 \leq w3 - w2 \leq 4.30$ (class 3, 2 false positives) (with the number of conjunctive terms restricted to be not larger than 2). Note that cluster 1 has been distinctively described already.

To further decrease the errors, after two more iterations of the procedure, we arrive at $0.64 \leq w2 * (w3 - w2) * w4 \leq 4.55$ & $0.21 \leq w2/(w3 * w4^2) \leq 0.74$ (class 2, 1 false positive) and $4.88 \leq w3 * w4^2 - w1 \leq 31.20$ & $-2.85 \leq (w3 - w2) * w4 - w1 \leq 2.19$ (class 3, 1 false positive). It should be added that class 1 can be distinctively separated with one of these variables, $w2 * (w3 - w2) * w4 \leq -3.07$ (class 1, 0 false positive).

The process of combining of the variables involves a trade-off between the exactness (number of false positives) and complexity of cluster descriptions (the complexity of combined variables and the number of conjunctive terms), which parallels similar trade-offs in other description techniques such as regression analysis.

6 Correspondence Analysis and Finding Associations

6.1 Box-Clustering

We refer to a nonnegative data matrix $P = (p_{ij})$, $i \in I$, $j \in J$, as a summable one if it makes sense to add the entries up to their total, $p_{++} = \sum_{i \in I} \sum_{j \in J} p_{ij}$, as it takes place for contingency, flow or mobility data.

There can be two different goals for the summable data analysis: 1) analysis between row and column set interrelations (this subsection), and 2) analysis within row and column similarities (next subsection). Both goals have been considered by researchers (for references, see Mirkin, Arabie and Hubert, 1995). Here, we show what approximation structuring can suggest for the goals.

To analyze row/column interrelations, a cluster structure called box clustering should be utilized. Two subsets, $V \subseteq I$ and $W \subseteq J$, and a real, μ, represent a box cluster as presented with $|I| \times |J|$ matrix having its entries equal to $\mu v_i w_j$ where v and w are Boolean indicators of the subsets V and W, respectively. This is how the admissible set D is defined here.

The question now is what data standardization option should be chosen so that the data reflect mutual dependence of the row and column items. Usually, conditional probabilities $p(s/t) = p_{st}/p_{+t}$, are considered as reflecting dependencies. However, better coefficients are available to compare the conditional probability $p(s/t)$ to the average rate p_s of s for all observations (see, for instance, Yule, 1900, p. 31). To make the comparison, the absolute change $w_{st} = p(s/t) - p_{s+}$,

or relative value $p(s/t)/p_{s+}$, or the relative change $q_{st} = (p(s/t) - p_{s+})/p_{s+}$ could be used. The relative value $p(s/t)/p_{s+} = p_{st}/(p_{s+}p_{+t})$ called the odds ratio, is a standard tool in contingency data analysis (see, for instance, Reynolds, 1977). The other two indices, w_{st} and q_{st}, have been suggested quite a while ago by Quetelet, 1832, as measures of the "degree of influence" of t towards s, as noted by Yule, 1900, p. 30-32. Of these two, the present author prefers the relative change index, q_{st}, because it shows no direction of influence (from t to s or from s to t) thus reflecting the postulate that no statistical data on its own can show the cause. All of the clustering and structuring contents of this section can be easily reformulated in terms of the other index, w_{st} (with corresponding changes in the results). Also, all the indices can be reformulated in general terms of summable flow data, with the (i, j)-th entry p_{ij} interpreted as amount of transaction from i to j. The ratio $p(j/i) = p_{ij}/p_{i+}$ shows the share of j in the total transactions of i, and $p(j/i)/p(j) = p_{ij}p_{++}/(p_{i+}p_{+j})$ compares the share of j in i's transactions with the share of j in the overall transactions, etc.

Table 7: Values of the Quetelet relative changes of probability (RCP), multiplied by 1000, for the Worries data.

	EUAM	IFEA	ASAF	IFAA	IFI
POL	222	280	-445	-260	36
MIL	168	-338	-129	-234	72
ECO	145	-81	-302	239	487
ENR	28	-40	11	-58	-294
SAB	19	-77	22	-66	-129
MTO	186	-252	-53	-327	-1000
PER	-503	-269	804	726	331
OTH	-118	582	-65	149	181

The data structuring model applied to matrix Q (with D consisting of box-cluster matrices $\alpha v w^T$) has the following form. The model is a bilinear equation,

$$q_{ij} = \sum_{t=1}^{m} \mu_t v_{it} w_{jt} + e_{ij} \tag{28}$$

to be fit by minimizing

$$L^2 = \sum_{i \in I} \sum_{j \in J} p_{i+}p_{+j}(q_{ij} - \sum_{t=1}^{m} \mu_t v_{it} w_{jt})^2 \tag{29}$$

with regard to real μ_t and Boolean v_{it}, w_{jt}, $t = 1, ..., m$.

Note that the least squares criterion is modified here: its (i, j)-th term is weighted with factor $p_{i+}p_{+j}$ which is important for further derivations.

It can be proven (see Mirkin, 1996b) that SEFIT, applied to model (28)–(29) without Boolean constraints with regard to v_{it}, w_{jt}, gives results coinciding with those of a well-known method in multivariate statistics, Correspondence Analysis (for a recent description of CA, see Lebart, Morineau and Piron, 1995).

In terms of the criterion (29), the data scatter is $\sum_{i \in I} \sum_{j \in J} p_{i+}p_{+j}q_{ij}^2$ which is just the Pearson chi-squared coefficient,

$$X^2 = \sum_{i \in I} \sum_{j \in J} \frac{(p_{ij} - p_{i+}p_{+j})^2}{p_{i+}p_{+j}}.$$

This implies that the data scatter decomposition according to the SEFIT criterion,

$$L^2 = \sum_{i \in I} \sum_{j \in J} p_{i+}p_{+j}(q_{ij} - \mu v_i w_j)^2,$$

is

$$X^2 = \mu^2 p_V p_W + L^2 \tag{30}$$

where μ is optimal, given the box $V \times W$. The optimal value of μ here is equal to

$$\mu(V, W) = \frac{p_{VW} - p_{V+}p_{+W}}{p_{V+}p_{+W}} \tag{31}$$

where $p_{VW} = \sum_{i \in V} \sum_{j \in W} p_{ij}$, and $p_{V+} = \sum_{i \in V} p_{i+}$, $p_{+W} = \sum_{j \in W} p_{+j}$. This means that the optimal intensity $\mu(V, W)$ is the same Quetelet coefficient $q = q_{VW}$, this time applied to subsets $V \subset I$ and $W \subset J$ rather than to items i and j. Equation (31) may be perceived as a major reason for using here the weighted least-squares fitting criterion (rather than nonweighted).

With (31) put in the contribution of box $V \times W$ into the data scatter according to (30), the contribution becomes

$$g(V, W) = \mu^2 p_{V+}p_{+W} = G^2(V, W)/p_{V+}p_{+W} \tag{32}$$

where

$$G(V, W) = \sum_{i \in V} \sum_{j \in W} p_{i+}p_{+j}q_{ij} \tag{33}$$

to be maximized by $V \subseteq I$ and $W \subseteq J$.

It is not known yet whether the problem of maximization (33) is NP-hard or not (though, it seems NP-hard). However, a local search method can be easily formulated based on a neighborhood system for boxes $V \times W$. Let us consider, in particular, such a neighborhood of $V \times W$ that includes every box of the form $V \times W'$ or $V' \times W$ where V' differs from V (and W' differs from W) by the addition or removal of only one element. With this neighborhood system, starting with empty V and W, the following local search algorithm for box clustering can be formulated.

Algorithm BOX
Start with $V = \{i\}$ and $W = \{j\}$ corresponding to maximum $f(\{i\}, \{j\}) = p_{i+}p_{+j}q_{ij}^2$ for $i \in I, j \in J$. Then, at any step, that one row i or column j is added to/removed from V or W, respectively, which maximizes the increment of $g(V, W)$ with respect to all $i \in I$ and $j \in J$. The process is finished when the maximum increment is not positive.

Although the algorithm above is quite simple, for box (V, W) found, the Quetelet relative change of probability (RCP) value within (V, W) deviates highly from the others.

Statement 6.1 *For any row i or column j outside the cluster box $V \times W$ found with the algorithm BOX, the absolute values of Quetelet coefficients, q_{Vj} and q_{iW}, are not greater than half the size of the absolute value of the integral Quetelet coefficient, q_{VW}, over the box.*

The proof of this statement can be found in Mirkin, 1996b. It shows that the data fragment corresponding to the box found reflects a pattern which is quite deviant from the general behavior. In the case when no such pattern exists, the algorithm BOX would lead to a box including all the rows and columns.
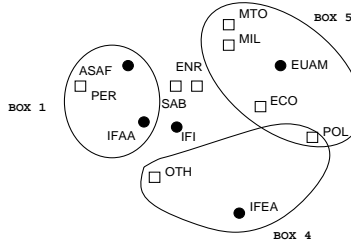
Figure 7: Positive flow boxes in the plane of the first two correspondence analysis factors.

Applied to the Worries data in Table 3, the algorithm BOX produces 6 clusters; the total contribution of the clusters in the initial value X^2 equals some 90 % (see Table 8).

Table 8: Box cluster structure of the Worries data set.

Box	Columns	Rows	RCP, %	Contrib., %
1	ASAF, IFAA	PER	79.5	34.5
2	EUAM, IFEA	PER	-46.0	20.8
3	ASAF, IFAA	POL, ECO	-40.5	9.9
4	IFEA	OTH, POL	46.1	9.7
5	EUAM	POL, MIL, ECO, MTO	18.5	9.3
6	IFEA, ASAF, IFAA, IFI	MIL, MTO	-17.5	5.5

The content of Table 8 corresponds to the traditional joint display given by the first two correspondence analysis factors (see Fig.7 where the columns and the rows are presented by the circles and the squares, respectively). Due to the model's properties, all the boxes with positive aggregate flow index (RCP, Quetelet coefficient) values (clusters 1, 4, and 5) correspond to the continuous fragments of the display (shown on Fig.7); boxes with the negative RCP values are associated with distant parts of the picture.

6.2   Aggregation of Flow Data

We refer to a box clustering problem as that of bipartitioning when the boxes are generated by partitions on each of the sets, I and J. Let $S = \{V_t\}$ be a partition of I, and $T = \{W_u\}$, of J, so that every pair (t, u) labels the corresponding box (V_t, W_u) and its weight μ_{tu}. In the corresponding specification of the model (28)-(29) for simultaneously partitioning the row and column sets, the optimal values μ_{tu} are the Quetelet coefficients $q_{V_t W_u}$ in (30).

Due to mutual orthogonality of the boxes (V_t, W_u), a decomposition of the weighted squared scatter of the data, q_{ij}, onto the minimized criterion L^2 (29) and the bipartition part which is just the sum of terms having the form of (32), can be made analogously to those above. An equivalent reformulation of the problem involves aggregation of the data based on the Pearson contingency coefficient. Let us aggregate the $|I| \times |J|$ table $P = (p_{ij})$ into $|S| \times |T|$ table $P(S, T) = (p_{tu})$ where $p_{tu} = \sum_{i \in V_t} \sum_{j \in W_u} p_{ij}$. In this notation, the original table is just $P = P(I, J)$. Then, the

contingency coefficient for $P(S,T)$ is

$$X^2(S,T) = \sum_{t,u} \frac{(p_{tu} - p_{t+}p_{+u})^2}{p_{t+}p_{+u}}.$$

It is not difficult to see, that the data scatter decomposition, due to the structuring model under consideration, is nothing but

$$X^2(I,J) = X^2(S,T) + L^2 \tag{34}$$

which means that the clustering and aggregation problems are equivalent in this setting:

Statement 6.2 *The bipartitioning problem according to the model (28)-(29) is equivalent to that of finding such an aggregate $P(S,T)$ which maximizes $X^2(S,T)$.*

Alternating and agglomerating optimization clustering procedures can be easily extended to this case (Mirkin, 1996b).
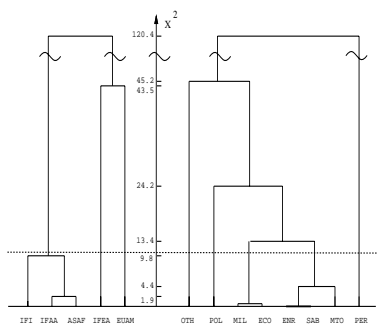

Figure 8: Hierarchical biclustering results for the Worries data: the vertical scale shows the level of decrement of the initial X^2 value at the correspondingly aggregated table.

6.3   Aggregation of Interaction Tables

The interaction table is a summable data table $P = (p_{ij})$ where $I = J$ as, for instance, in brand switching or digit confusion or input-output tables. The aggregation problem for such a table can be stated as that of bipartitioning with coinciding partitions, $S = T$, or equivalently, of finding an aggregate table $P(S,S)$ maximizing the corresponding Pearson contingency coefficient $X^2(S,S)$. Another formulation involves finding such a partition, $S = \{V_1, ..., V_m\}$, that the aggregate flow index values (Quetelet coefficients), q_{tu}, satisfy equations

$$q_{ij} = q_{tu} + \epsilon_{ij}, \ i \in V_t, \ j \in V_u \tag{35}$$

and minimize the criterion, $\sum_{t,u} \sum_{i \in V_t} \sum_{j \in V_u} p_{i+}p_{+j}(q_{ij} - q_{tu})^2$.

Applying the agglomerative clustering algorithm (by minimizing the decrement of $X^2(S,S)$ at each agglomeration step) to the Confusion data table (all the entries taken into account), we obtain the hierarchy presented in Fig. 9. The hierarchy is indexed by the level of unexplained X^2 at each level of aggregation.

The aggregate confusion rate and Quetelet coefficient data corresponding to the four-class partition, $S = \{\{1,4,7\}, \{3,5,9\}, \{6,8,0\}, \{2\}\}$, is in Table 9:
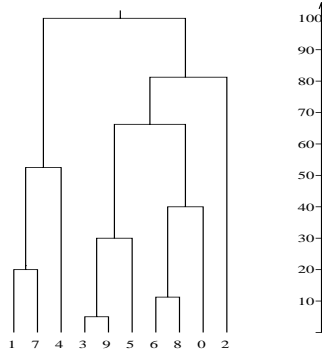
Figure 9: Results of agglomerative chi-square based aggregation for the Confusion data.

Table 9: **Confusion:** Original aggregate data and corresponding Quetelet coefficients.

I	2827	33	96	44	I	1.95	-0.88	-0.89	-0.95
II	33	783	90	94	II	-0.90	7.29	-0.69	-0.68
III	203	85	2432	280	III	-0.79	-0.70	1.81	-0.69
IV	134	46	263	2557	IV	-0.84	-0.84	-0.70	1.86

7 Discrete Hierarchies, Quadtrees and Wavelets

7.1 Multiresolution Approximation via Binary Hierarchies

In this section, only resolved binary hierarchies will be considered, having both I and all singletons, $\{i\}$ ($i \in I$), belonging to S_W so that $E = 0$ in equation (12). The concept of binary hierarchy fits into spatial data structures: digitized intervals, rectangles or hyper-rectangles consisting of one-, two- or three- dimensional pixels arranged in grids according to the coordinate axes (Samet, 1990). Let us consider initially I to be a unidimensional pixel set.

In problems of data compression, the hierarchy layers (which are obtained by cutting the tree at any level) can be exploited for approximate compression of the data. More specifically, with a layer $L_m = \{L_{mt}\}$ taken, a data vector $f = (f_i)$, $i \in I$, can be substituted by the vector of within class averages, $f_{mt} = \sum_{i \in L_{mt}} f_i / |L_{mt}|$, which is considered as the data at the m-th level of resolution. The smaller m, the coarser the resolution; the larger m, the finer the resolution.

The layers can be trivially used for recalculating the averages while running along the hierarchy bottom-up. It is not difficult also to exploit the hierarchy for recalculating the averages running up-down along the hierarchy. Let us save, for every cluster S_w, in addition to f_w, the between-split difference $d_w = f_{w1} - f_{w2}$. The formulas

$$f_{w1} = f_w + \frac{n_{w2}}{n_w} d_w, \quad f_{w2} = f_w - \frac{n_{w1}}{n_w} d_w \tag{36}$$

provide for calculating the average values in L_{m+1} by the averages of L_m. This allows one to carry out the decompression of the data quickly.

In Fig. 10, two hierarchies, A and B, are exploited for compressing a vector f whose values
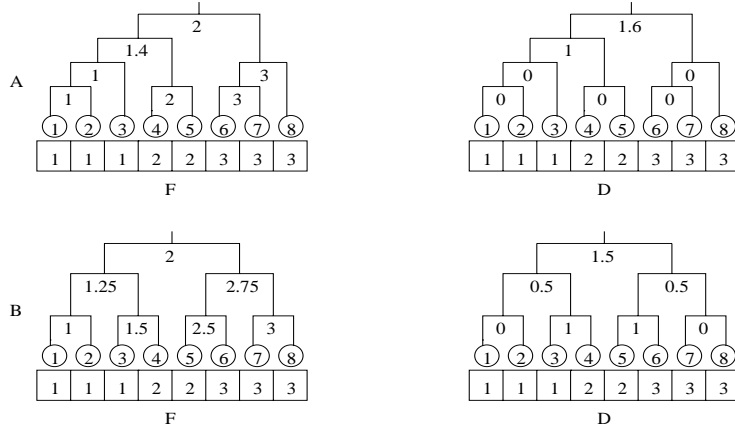
31

Figure 10: Compression and decompression of the boxed data with two hierarchies, A and B. The latter is Haar wavelet hierarchy from Fig. 3.

are the boxed digits: version F keeps all the averages, D all the differences. It can be seen that hierarchy A provides for a safer data compression: only one average, f_0 in F, and two differences, 1.6 and 1, are needed to decompress the data entirely. Thus, adjusting the hierarchy to the data may lead to better processing.

This can be put in the linear space framework as follows.

Let us define, for the partition at any layer L_m, corresponding binary matrix of membership vectors; its columns are a basis of the corresponding subspace, V_m. Let us denote by D_m the subspace generated by the nest indicator vectors, $\phi_{mt}(i)$, of nonsingleton classes in L_m. It appears, for any m $(m = 1, ..., q)$, subspace D_{m-1} is the orthogonal complement of V_{m-1} in V_m so that $V_{m-1} \oplus D_{m-1} = V_m$.

This can be "decoded" into interconnection between coefficients of decompositions of the vector f through subspaces of different levels. It appears that the equations above - in the case of a complete binary hierarchy (as B in Fig. 10) - are exactly parallel to those emerging in the theory for multiresolution approximation involving Haar basis (Mirkin, 1997b).

7.2 Finding and Using Quadtrees for Data Processing

These unidimensional constructions can be extended onto two-dimensional pixellated images via the following concept. A hierarchy S_W defined on $I = I' \times I''$ will be referred to as a *bihierarchy* if any of its clusters S_w is a Cartesian rectangle, that is, $S_w = A \times B$ for some $A \subseteq I'$ and $B \subseteq I''$, and the children of S_w are $A1 \times B1$, $A1 \times B2$, $A2 \times B1$, and $A2 \times B2$ for some partitions, $\{A1, A2\}$ and $\{B1, B2\}$, of A and B, respectively. (To allow more freedom in handling "one-dimensional" strip clusters, $\{i'\} \times B$ or $A \times \{i''\}$, we can admit some of the subsets as being empty.) The sets, A and B, can be referred to as the ranges of S_w in I' and I'', respectively. A bihierarchy will be called *spatial* if I' and I'' are ordered and the ranges of all clusters are intervals of these orders. A specific case of a bihierarchy is the Cartesian product of two binary hierarchies, $S_W = S'_{W'} \times S''_{W''}$, the clusters of which are all possible Cartesian products of clusters of $S'_{W'}$ and $S''_{W''}$.

A (divisive) bihierarchical cluster structure is an "upper" part of a bihierarchy.
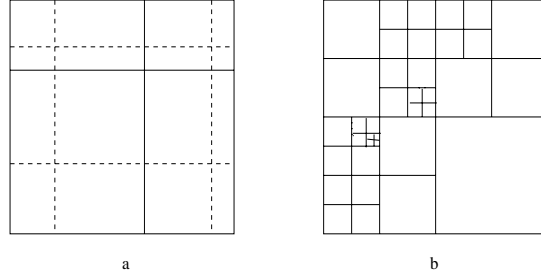

Figure 11: Higher splits of a Cartesian product of two spatial binary hierarchies (a) and a quad-tree (b).

A well-known structure in image data analysis, the quadtree (see, for example, Samet, 1990) fits into this: a quadtree is just a bihierarchical cluster structure for a complete spatial bihierarchy (see Fig. 11, (b)).

For a cluster S_w in a bihierarchy, S_W, with its ranges A and B subdivided into $A1, A2$ and $B1, B2$, respectively, three nest indicator functions are needed to linearly represent its four children. A natural way of defining the indicators would be by considering the four children as produced within a binary hierarchy via double dichotomy. In such a double dichotomy cluster $S_w = A \times B$ can be divided, firstly, into two strips, say, $A1 \times B$ and $A2 \times B$, and secondly, each of the strips is further split into the final children $Ak \times Bj$, $k, j = 1, 2$. The three splits can be assigned with corresponding nest indicator functions. The bihierarchy can be regarded as a contracted version of the binary hierarchy involving the double dichotomy described.

For the sake of computational simplicity, we consider here yet another triple of nest indicator functions. Each of the ranges implies its nest indicator function, $\phi_A(i')$ and $\phi_B(i'')$. The three cluster nest indicator functions, ϕ_A, ϕ_B, and ϕ_{AB}, then, can be defined for all $(i', i'') \in S_w = A \times B$ as (1) $\phi_A(i', i'') = \phi_A(i')\chi_B(i'')$, (2) $\phi_B(i', i'') = \chi_A(i')\phi_B(i'')$, and (3) $\phi_{AB}(i', i'') = \phi_A(i')\phi_B(i'')$ where $\chi_S(i) = 1/\sqrt{(|S|)}$ when $i \in S$ and $= 0$ when $i \notin S$. (When A or B is a singleton, only one of these three functions remains valid.) These functions, obviously, are centered and normed (with regard to all $(i', i'') \in I' \times I''$) and, moreover, are mutually orthogonal. Thus, the nest indicator functions of all interior clusters $S_w \in S_W$ form an orthonormal basis, Φ, of the space of $|I' \times I''|$-dimensional centered matrices (considered as vectors). The coefficients of decomposition of a matrix vector $y(i', i'')$ defined on $I' \times I''$ by the fragment of Φ related to a cluster $S_w = S_{AB}$ are scalar products of $y(i', i'')$ and corresponding nest indicator functions that can be shown to have the following form:

$$c_A = \sqrt{\frac{n_{A1}n_{A2}}{n_A}}\sqrt{n_B}(y_{1.} - y_{2.}),$$

$$c_B = \sqrt{n_A}\sqrt{\frac{n_{B1}n_{B2}}{n_B}}(y_{.1} - y_{.2}) \tag{37}$$

$$c_{AB} = \sqrt{\frac{n_{A1}n_{A2}}{n_A}}\sqrt{\frac{n_{B1}n_{B2}}{n_B}}(y_{11} - y_{12} - y_{21} + y_{22})$$

where y_{kj}, $y_{k.}$, or $y_{.j}$ is the average of $y(i', i'')$ on $Ak \times Bj$, $Ak \times B$ or $A \times Bj$, respectively $(k, j = 1, 2)$.

These expressions can be easily extended to the situation of three-way data $Y = (y(i', i'', k))$ by adding an index k where necessary.

Usually, quadtrees are utilized for storing images only. In the framework presented, two more developments can be suggested: clustering and compression/decompression of data.

Let us discuss clustering first. For clustering, we need to relax conditions of continuity and equality of subdivisions in quadtrees, which is quite easy in terms of bihierarchies. Following the sequential extraction SEFIT strategy, we arrive at the problem of splitting the ranges of a given rectangle $A \times B \subseteq I' \times I''$ to maximize $\mu_{AB}^2 = c_A^2 + c_B^2 + c_{AB}^2$ where the items are defined in (37):

$$\mu_{AB}^2 = \frac{n_{A1}n_{A2}}{n_A}\frac{n_{B1}n_{B2}}{n_B}(y_{11} - y_{12} - y_{21} + y_{22})^2 + \frac{n_{A1}n_{A2}}{n_A}n_B(y_{1.} - y_{2.})^2 + n_A\frac{n_{B1}n_{B2}}{n_B}(y_{.1} - y_{.2})^2 \quad (38)$$

This can be done with a local search algorithm. For instance, to find an initial partition, let us split A to maximize c_A^2 and, in parallel, B to maximize c_B^2. This can be done with an algorithm for splitting a cluster described in section 3.2. Then, the partition found can be iteratively updated by exchanging rows between $A1$ and $A2$ or columns between $B1$ and $B2$ (one item in a time) until μ_{AB}^2 cannot be increased anymore.

The issue of plane image data compression and decompression can be considered in the same fashion as described above for hierarchies. We will not maintain here the linear subspace terminology since it does not much differ from that described above. Let us just show how data compressed as within cluster averages can be decompressed up-down employing the three differences involved in (37) and kept as coefficients of the "wavelet" bases consisting of those parts of Φ that correspond to layers of a bihierarchy S_W:

$$d_{AB} = y_{11} - y_{12} - y_{21} + y_{22}, \ d_A = y_{1.} - y_{2.}, \ d_B = y_{.1} - y_{.2}.$$

Statement 7.1 *In a bihierarchy, the children's averages can be expressed through the within cluster S_w average, y_w, and the d-coefficients above as follows:*

$$y_{11} = y_w + \frac{n_{A2}}{n_A}\frac{n_{B2}}{n_B}d_{AB} + \frac{n_{A2}}{n_A}d_A + \frac{n_{B2}}{n_B}d_B,$$

$$y_{12} = y_w - \frac{n_{A2}}{n_A}\frac{n_{B1}}{n_B}d_{AB} + \frac{n_{A2}}{n_A}d_A - \frac{n_{B1}}{n_B}d_B,$$

$$y_{21} = y_w - \frac{n_{A1}}{n_A}\frac{n_{B2}}{n_B}d_{AB} - \frac{n_{A1}}{n_A}d_A + \frac{n_{B2}}{n_B}d_B,$$

$$y_{22} = y_w + \frac{n_{A1}}{n_A}\frac{n_{B1}}{n_B}d_{AB} - \frac{n_{A1}}{n_A}d_A - \frac{n_{B1}}{n_B}d_B.$$

Proof: The proof follows with a little arithmetic from the basic equations connecting y_w, $y_{k.}$ and $y_{.j}$ with y_{kj}, $k, j = 1, 2$, as, for instance $n_A n_B y_w = n_{A1}n_{B1}y_{11} + n_{A1}n_{B2}y_{12} + n_{A2}n_{B1}y_{21} + n_{A2}n_{B2}y_{22}$, and definitions of d_{AB}, d_A, d_B. $\square$

These formulas can be converted into the language of V_m and D_m spaces as was done in the case of hierarchies.

8 Conclusion

The results presented show that the framework of approximation clustering amounts to a mathematical theory that not only meets some direct theoretical and computational clustering needs, but also establishes firm connections of clustering with seemingly unrelated methods and problems. Although some of the answers proposed to the issues raised may not seem decisive, they have nice properties derived in the least-squares context. Also, a number of interpretation aids emerging in the approximation framework have been presented, especially in sections 4.2, 4.3, 4.4, and 5.2.

References

[1] H.H. Bock (1974) *Automatische Klassifikation*, Goettingen: Vandenhoeck and Ruprecht.

[2] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone (1984) *Classification and Regression Trees*, Belmont, Ca: Wadswarth International Group.

[3] A. Chaturvedi and J.D. Carroll (1994) An alternating optimization approach to fitting INDCLUS and generalized INDCLUS models, *Journal of Classification, 11*, 155-170.

[4] A.W.F. Edwards and L.L. Cavalli-Sforza (1965) A method for cluster analysis, *Biometrics, 21*, 362-375.

[5] U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth (1996) From data mining to knowledge discovery: An overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, Menlo Park, Ca: AAAI Press/The MIT Press, 1-37.

[6] D.H. Fisher (1987) Knowledge acquisition via incremental conceptual clustering, *Machine Learning, 2*, 139-172.

[7] M.J. Greenacre (1988) Clustering the rows and columns of a contingency table, *Journal of Classification, 5*, 39-51.

[8] L. Guttman (1971) Measurement as structural theory, *Psychometrika, 36*, 329-347.

[9] L. Hubert and P. Arabie (1994) The analysis of proximity matrices through sums of matrices having (anti)-Robinson forms, *British Journal of Mathematical and Statistical Psychology, 47*, 1-40.

[10] L. Hubert and P. Arabie (1995) Iterative projection strategies for the least-squares fitting of tree structures to proximity data, *British Journal of Mathematical and Statistical Psychology, 48*, 281-317.

[11] A.K. Jain and R.C. Dubes (1988) *Algorithms for Clustering Data*, Englewood Cliffs, NJ: Prentice Hall.

[12] J. Kay (1994) Wavelets, *Advances in Applied Statistics, 2*, 209-224.

[13] G. Keren and S. Baggen (1981) Recognition models of alphanumeric characters, *Perception and Psychophysics, 29*, 234-246.

[14] L. Lebart, A. Morineau and M. Piron (1995) *Statistique Exploratoire Multidimensionnelle*, Paris: Dunod.

[15] R.S. Michalski and R.E. Stepp (1992) Clustering. In S.C. Shapiro (Ed.), *Encyclopedia of artificial intelligence*, New York: J. Wiley and Sons.

[16] B. Mirkin (1990) A sequential fitting procedure for linear data analysis models, *Journal of Classification, 7*, 167-195.

[17] B. Mirkin (1996a) *Mathematical Classification and Clustering*, Dordrecht: Kluwer Academic Press.

[18] B. Mirkin (1996b) Clustering for contingency tables: boxes and partitions, *Statistics and Computing, 6*, 217-229.

[19] B. Mirkin (1997a) L_1 and L_2 approximation clustering for mixed data: scatter decompositions and algorithms. In Y. Dodge (Ed.) L_1-*Statistical Procedures and Related Topics*, Institute of Mathematical Statistics: Lecture Notes-Monograph Series, Hayward, Ca., 473-486.

[20] B. Mirkin (1997b) Linear embedding of binary hierarchies and its applications. In B. Mirkin, F. McMorris, F. Roberts, A. Rzhetsky (Eds.) *Mathematical Hierarchies and Biology*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Providence, RI: AMS.

[21] B. Mirkin (1998) Concept learning and feature selection based on square-error clustering, *Machine Learning*, to appear.

[22] B. Mirkin, P. Arabie and L Hubert (1995) Additive two-mode clustering: the error-variance approach revisited, *Journal of Classification, 12*, 243-263.

[23] G. Pagallo and D. Haussler (1990) Boolean feature discovery in empirical learning. *Machine Learning, 5*, 71-99.

[24] A. Quetelet (1832) Sur la possibilité de mesurer l'influence des causes qui modifient les éléments sociaux, *Lettre à M. Willermé de l'Institut de France*, Bruxelles.

[25] H.T. Reynolds (1977) *The Analysis of Cross–Classifications*, New York: The Free Press.

[26] H. Samet (1990) *The Design and Analysis of Spatial Data Structures*, Addison-Wesley Series on Computer Science and Information Processing. Addison-Wesley Publishing Company.

[27] R.N. Shepard and P. Arabie (1979) Additive clustering: representation of similarities as combinations of overlapping properties, *Psychological Review, 86*, 87-123.

[28] P.H.A. Sneath and R.R. Sokal (1973) *Numerical Taxonomy*, San Francisco: W.H. Freeman.

[29] J.H. Ward, Jr (1963) Hierarchical grouping to optimize an objective function, *Journal of American Statist. Assoc., 58*, 236-244.

[30] J. Wnek and R.S. Michalski (1994) Hypothesis-driven constructive induction in AQ17-HCI: A method and experiments. *Machine Learning, 14*, 139-168.

[31] G.U. Yule (1900) On the association of attributes in statistics: with illustrations from the material of the Childhood Society, *Phil. Trans., A, 194*, 257-319. In A. Stuart and M.G. Kendall (1971) *Statistical Papers of George Udny Yule*, New York: Hafner Publishing Company, 7-70.